Procedia Computer Science

Volume 80, 2016, Pages 2397–2401

ICCS 2016. The International Conference on Computational Science

Detecting Extreme Events in Gridded Climate Data

Bharathkumar Ramachandra[1], Krishna Karthik Gadiraju[1], Ranga Raju Vatsavai[1], Dale P. Kaiser[2], and Thomas P. Karnowski[2]

[1] North Carolina State University, Raleigh, North Carolina, U.S.A.
bramach2@ncsu.edu, kgadira@ncsu.edu, rrvatsav@ncsu.edu
[2] Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A.
kaiserdp@ornl.gov, karnowskitp@ornl.gov

Abstract

Detecting and tracking extreme events in gridded climatological data is a challenging problem on several fronts: algorithms, scalability, and I/O. Successful detection of these events will give climate scientists an alternate view of the behavior of different climatological variables, leading to enhanced scientific understanding of the impacts of events such as heat and cold waves, and on a larger scale, the El Niño Southern Oscillation. Recent advances in computing power and research in data sciences enabled us to look at this problem with a different perspective from what was previously possible. In this paper we present our computationally efficient algorithms for anomalous cluster detection on climate change big data. We provide results on detection and tracking of surface temperature and geopotential height anomalies, a trend analysis, and a study of relationships between the variables. We also identify the limitations of our approaches, future directions for research and alternate approaches.

Keywords: spatio-temporal, co-location, anomaly detection, trend analysis

1 Introduction

Extreme climatic events include phenomena such as heat waves, cold waves, floods and cyclones. According to [11], "it is very likely that frequency of heat waves has increased in large parts of Europe, Asia and Australia". To predict the occurrence of extreme events and mitigate their impact on economy and human life, there is an increasing need to study climate data. Techniques to find anomalous behavior are often called anomaly detection.

1.1 Problem Description

Ways to detect clusters of similar behavior include density and correlation-based techniques among others [6] [7]. Some of these methods can be extended to be applied on data of spatial and/or temporal nature [10]. One of the open problems in this domain is to detect and track extreme events through space and time.

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.537

The task is made harder by spatial and temporal context introduced by locality. Figure 1 represents data for one day, comprised of 10,512 discrete spatial samples. Considering a history of 35 years approximately comes to 134 million values. These values are from the sampling of NCEP Reanalysis data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA (http://www.esrl.noaa.gov/psd/) [4]. The data are represented in a 73x144 gridded format (2.5 by 2.5 latitude longitude). The computational and I/O challenges are huge as the spatio-temporal methods used are more complex.

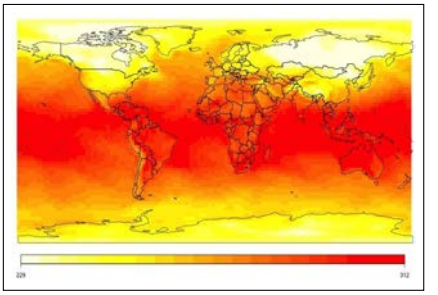

Figure 1: Surface temperatures across the Earth's surface on Jan 1st 2010. Temperatures are in Kelvin.

An appropriate term to define the problem is *cluster detection*. Cluster detection is nicely stated in Neill's thesis [9]; in clustering the number of clusters is almost always previously specified whereas in cluster detection, whether there exist anomalous clusters or not is in itself an important question. Anomaly detection focuses on point anomalies but cluster detection searches for collective (grouped point) anomalies. A set of similarly hot anomalous values that are spatially close could be indicative of a heat wave. The fact that we are interested in groups of anomalies rather than point anomalies poses an additional computational challenge. Point anomalies might be coincidences or artifacts that were introduced during data preprocessing [2] but this is less likely for spatially and temporally contiguous (or groups of) anomalies.

1.2 Our Contributions

The main idea of our work is to use statistical learning techniques to build a pipeline for analysis of climate extremes. We have chosen to work with *"height of the 500 millibar atmospheric pressure level"* (which will henceforth be referred to as Z500) and *"near-surface temperature"* (which will henceforth be referred to as Tsfc). The Z500 level is thought to be higher in the atmosphere when the temperature in the column below is relatively higher, causing the air to expand. Similarly, a low Z500 level is thought to correspond to colder air in the column below. The computational approach that we developed leverages the concept of pure temporal analysis in the first step while ignoring spatial autocorrelation.

2 Methods Used

Our approach uses a purely temporal analysis documented in [8] to detect point anomalies. This is followed by a connected components analysis to detect the grouped anomalies. Consequently, we mine information about these anomalous regions in order to better characterize their properties.

2.1 Purely Temporal Analysis

The concept of temporal neighborhood is utilized; the value of a variable at a particular location on a given day is candidate for comparison with values at the same location on days immediately preceding and following it and the same window of days of other years. This is justified as we are concerned primarily with how the tails of the distribution change over space and time, not with change around the mean[5].

We considered a 5-day window centered on a calendar day. Over 35 years, this yields a set of 174 values (excluding the one under observation) for comparison for each day of the year. The mean of these values is considered the long-term mean for this particular location. We say that the location on the day observed experiences anomalous behavior if the temperature lies in the tails of these 175 values. The anomaly score for each location and day of the year is computed by subtracting its long-term mean. We used different percentile thresholds ranging in {10, 90}, {5, 95} and {1, 99} in order to provide a richer picture of behavior. Two-tailed percentiles were always used as both hot and cold anomalous events are of interest to us.

2.2 Connected Components Analysis

Connected components analysis [1] is a well known image processing technique to identify groups distinctly. We create an anomaly mask, then iterate through the gridded data row wise and examine each grid cell's immediate eight surrounding neighbors for contiguity and stop when a stable grouping is confirmed.

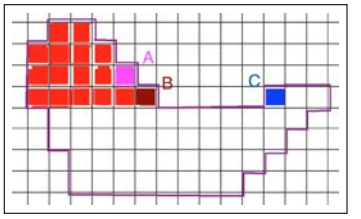

Figure 2: Computational challenges with connected component analysis.

Since our way to iterate through the data is restricted to a fixed order (row wise), a group of anomalies could need many iterations of the data before we can conclude that we have finalized the grouping. Figure 2 aims to demonstrate this computational challenge. As shown in the figure, the purple outline represents the contiguous anomalous group to be discovered, say a heat wave. For the case of point A, we have no problem recognizing the values on this row as participants of the anomalous region. For point B, clearly its left neighbor is anomalous, but it is not possible to link to point C without analysis of subsequent rows.

2.3 Extracting Properties of Anomalous Regions

To capture properties of anomalous regions, we identified four essential properties - *surface area*, *magnitude*, *duration* and *frequency*. The surface area and magnitude properties are relatively straightforward to compute. Mining frequency is also a straightforward counting procedure that follows naturally from Section 2.2. To track regions through space and time to infer duration, we used an algorithm based on the simple idea of "surface area overlap". If a certain percentage of surface area enclosed in anomalous region A on day i is also present in an anomalous region B on day i + 1, we say A and B are part of the same "extreme event". Overview of key details:

- Colors are used analogous to unique identifiers on a given day, so tracing a color studies the duration of an extreme event.
- Assume a maximum number of cold and hot regions on a given day (conservative estimate) in order to prevent unbounded usage of number (color) space.
- Regions can be of any shape and size.
- An anomalous region may have only one successor but a successor may have more than one predecessor.

3 Uncovering Relationships

We employed a combination of trend analysis and co-incidence mining to find relationships between the variables.

3.1 Trend Analysis

Property		Z500				Tsfc			
	Tail	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
Magnitude (Z, Temp.)	Cold	+	+	+	+	+	+	+	+
	Warm	0	0	+	+	0	0	+	+
Frequency	Cold	-	-	-	-	-	-	-	-
	Warm	+	+	+	+	+	+	+	+
Surface Area	Cold	-	0	-	-	-	0	-	-
	Warm	0	+	+	+	0	+	+	+
Duration	Cold	0	0	0	0	0	0	0	0
	Warm	0	+	+	0	0	0	+	0

Table 1: Seasonal trends for four anomalous region properties of Z500 and Tsfc: magnitude, frequency, surface area and duration. Sign and significance are shown (gray boxes; p-value < 0.05) and 0 denotes no significant linear trend.

We aggregated the properties found in Section 2.3 annually to perform a trend analysis. Segmenting the trend analysis into the 4 seasons of the year corresponding to winter (Dec/-Jan/Feb), spring (MAM), summer (JJA) and autumn (SON) is common practice as the behavior of extremes is known to differ greatly among different seasons. T-tests were performed at the 95% significance level to test for the significance of linear relationships between each property and time. The results are summarized in Table 1 and observations from it are:

- Magnitude of cold tail events has "warmed" for both variables across all seasons.
- Magnitude of Z500 warm tail events has only increased in summer and fall.
- Fewer (more) cold (warm) tail events for both Z500 and Tsfc across all seasons.

3.2 Co-incidence Patterns

We studied co-location on a fine-grained scale by creating a heat map of co-location fractions for each grid cell. Figure 3 shows the heat maps. Each grid cell has the number of times both Tsfc and Z500 extremes occurred divided by the number of times only a Z500 extreme occurred. We observe that extreme events are more "stacked" in summer, corresponding to the northward retreat of the jet stream, less amplified ridges and troughs compared to winter, and a more "equivalent barotropic" pattern.

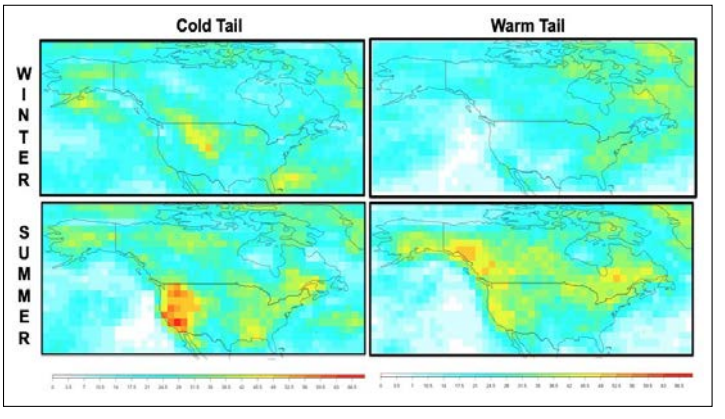

Figure 3: Percentage of days co-location was observed.

4 Conclusions and Future Work

In this work, we created a computational pipeline to study extreme events in historical climate change big data. Having this modular pipeline means that for future analysis we can swap the individual components with better approaches. Our analysis produced results that were not only in agreement with previously published research results, but also novel. Stacking of extreme events in the summer over the western part of North America is especially interesting.

Initial analysis validates our results with extreme events recorded in history [3]. However, these methods do not account for spatial autocorrelation as they consider each grid cell separately, without accounting for neighbors. Our accuracy was compromised a little as we excluded the first two and last two days of every year in considering the 5 day windows to save some computation. We also lost a little information as the neighborhood for locations along a grid's edge were excluded from the analysis to also save some compute time. The co-incidence mining methodology can be improved by accounting for spatial and temporal lags in addition to studying co-locations between events rather than each grid cell separately.

References

[1] Michael B. Dillencourt, Hanan Samet, and Markku Tamminen. A General Approach to Connected-component Labeling for Arbitrary Image Representations. *J. ACM*, 39(2):253–280, April 1992.

[2] Markus G. Donat, Jana Sillmann, Simon Wild, Lisa V. Alexander, Tanya Lippmann, and Francis W. Zwiers. Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets*. *Journal of Climate*, 27(13):5019–5035, 2014.

[3] Martin Hoerling, Arun Kumar, Randall Dole, John W. Nielsen-Gammon, Jon Eischeid, Judith Perlwitz, Xiao-Wei Quan, Tao Zhang, Philip Pegion, and Mingyue Chen. Anatomy of an Extreme Event. *Journal of Climate*, 26(9):2811–2832, November 2012.

[4] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, Roy Jenne, and Dennis Joseph. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77(3):437–471, March 1996.

[5] Evan Kodra and Auroop R. Ganguly. Asymmetry of projected increases in extreme temperature distributions. *Scientific Reports*, 4, July 2014.

[6] Hans-Peter Kriegel, Peer Krger, Jrg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, May 2011.

[7] Hans-Peter Kriegel, Peer Krger, and Arthur Zimek. Clustering High-dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering. *ACM Trans. Knowl. Discov. Data*, 3(1):1:1–1:58, March 2009.

[8] Paul C. Loikith and Anthony J. Broccoli. Characteristics of Observed Atmospheric Circulation Patterns Associated with Temperature Extremes over North America. *Journal of Climate*, 25(20):7266–7281, October 2012.

[9] Daniel B Neill. *Detection of spatial and spatio-temporal clusters*. PhD thesis, University of South Carolina, 2006.

[10] R.T. Ng and Jiawei Han. CLARANS: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):1003–1016, September 2002.

[11] Rajendra K Pachauri, MR Allen, VR Barros, J Broome, W Cramer, R Christ, JA Church, L Clarke, Q Dahe, P Dasgupta, et al. Climate change 2014: Synthesis report. contribution of working groups i, ii and iii to the fifth assessment report of the intergovernmental panel on climate change. 2014.